On Knowing What's Right and Being Responsible for it

---

1. Introduction

Does an agent's epistemic situation affect what she is morally responsible for? My aim in this paper is to defend a positive answer to this question. I argue that moral responsibility is closely related to an agent's *moral knowledge*. Moral responsibility, as I will understand it here, has to do with the appropriateness of certain reactive attitudes, such as praise and blame. Thus, to say that an agent is morally responsible for an action is to say that there is some reactive attitude that it's appropriate to take towards her in response to her action.

My focus here will be on moral responsibility for *right actions*. Even if we lack a term for it, there clearly is a reactive attitude associated with giving people credit for having done good. Here, I agree with Gary Watson:

> To be held liable is to be on the hook, and we lack a ready phrase for the positive counterpart to the "hook." But clearly we do have a counterpart notion; just as (moral) blame is sometimes called for as a response to the flouting of (moral) requirements, so praise is an appropriate response to respect for moral requirements or moral ends. We express praise by recognition: bestowing a medal, or, more commonly, remarking on the person's merits. ("It was good of you [him] to help.")[1]

The central claim in this paper is that whether an agent is morally responsible for her right action depends on whether she knows what the right thing to do is. Moral knowledge matters to moral evaluations because it's a central ingredient in intentional action. Our knowledge of what the right and wrong thing to do is, in part, what determines whether we do the right or wrong thing intentionally. Moral responsibility inherits its epistemic condition from the epistemic condition on intentional action.

Let me give you a preview of my argument:

> *Premise 1*: Intentional action involves an epistemic condition: an agent intentionally does what's right only if she knows what the right thing to do is.
> *Premise 2*: An agent is morally responsible for doing what's right only if she intentionally does what's right.
> *Conclusion*: And so, an agent is morally responsible for a right action only if she knows what the right thing to do is.

Premise 1 makes a general claim about the nature of intentional action and I outline the motivations for it in Section 2. But my main concern in this paper is Premise 2 – a claim about the moral psychology of morally praiseworthy actions. The main aim of this paper is to argue in support of this premise: to motivate it, to defend it against alleged counterexamples, and to make a case that we should prefer it over alternative accounts of moral worth. Sections 3-5 are concerned with just that, in that order. Section 6 draws out an implication of the present account for the limits of moral responsibility.

2. Intentionally Doing the Right Thing

I will start by setting out an account of intentionally doing the right thing. I'm going to proceed in two steps. First, I argue that to intentionally perform an action, an agent needs to know how to perform the action. Second, I distinguish between intentionally performing an action that is right and intentionally doing the right

---

[1] Watson [2004], p. 284.

thing. While an agent need not know the moral status of her action for the former, intentionally doing the right thing does require such knowledge.

At any given moment, unless we are asleep, we perform a number of actions. I write a paragraph. I press various keys on my keyboard. I let my gaze sweep around the room, looking for distractions. I take breaths. Some of these things I do intentionally (writing a paragraph), others I do not do intentionally (taking breaths). What distinguishes intentional from unintentional action? The term "intentional action" strongly suggests one answer: what distinguishes intentional from unintentional actions is that the former but not the latter are appropriately related to an intention. I have an intention to write a paragraph and my finger movements across the keyboard can be traced to this intention. But I do not have an intention to take breaths. (This is not to say that I couldn't breath intentionally – just that, as a matter of fact at this moment, I don't.) Unsurprisingly then, a central theme in the philosophy of action has been to give an account of intentions and how they figure in intentional action.[2] This is an important debate but we shall leave it aside. What matters for the purposes of this paper is that whether an action is intentional also depends on the agent's epistemic circumstances.

We can bring this out by an example: I intend to participate in a lottery, hoping to win. I buy a lottery ticket. It turns out to be the winning ticket. Which of my actions have been intentional? Plausibly, I have intentionally bought *a* lottery ticket. But have I intentionally bought the *winning* lottery ticket? That seems like a stretch. Perhaps, I could have intended to buy the winning ticket. But since I had no idea that the ticket I was about to buy was the winning ticket, I couldn't have bought the winning ticket intentionally. This suggests that there is some epistemic condition on intentional action: an agent has not performed an action intentionally unless she satisfies the condition in question.

A natural idea is that in order to intentionally perform an action, I need to know how to perform it. I couldn't have intentionally bought the winning lottery ticket because I didn't know which ticket was going to win the lottery; and so, I didn't know how to win the lottery.[3] The competence requirement is motivated by the general observation that whether we are inclined to judge an action as intentional or not seems to be systematically sensitive to considerations of luck. The kind of luck that's incompatible with an action's being intentional, seems to be the same kind of luck that intuitively undermines an agent's knowledge of how to perform the action.[4] To say that someone acted intentionally implies not just that their action was successful but also that there was something about how the action was motivated that made their success counterfactually robust. If the agent's action was guided by her knowledge of how to perform it, this counterfactual robustness is easy to explain. Knowledge, after all, is by its nature counterfactually robust; it requires safety.[5]

---

[2] See Bratman [1987], Mele [1992] for the former, Velleman [2000], Setiya [2008] for the latter.

[3] We can steer clear of issues about whether knowledge how reduces to propositional knowledge. What matters for my purposes is that sometimes an agent can lack knowledge how because she lacks propositional knowledge.

[4] See Mele [1994] for a discussion of various cases. Mele concludes that there is some epistemic requirement for intentional action but does not endorse the specific principle that I am putting forward.

[5] Again, this is independent of the question whether knowledge how reduces to propositional knowledge. See Hawley [2003] for a discussion of counterfactual robustness of know how.

Taking intentional action to require know how has other theoretical payoffs. Analyses of intentional action in terms of an agent's beliefs, desires, and other necessary conditions have run a similar course to the post-Gettier project of giving a reductive analysis of knowledge in terms of belief, truth, and other conditions. If intentional action requires knowledge and knowledge is, as Williamson has argued, unanalyzable, then it's not surprising that a viable reductive analysis of intentional action in terms of belief and further conditions has not been forthcoming.[6]

So far, my focus has been on intentional action in general. I have argued that to intentionally perform an action, an agent needs to know how to perform it. Let us now turn to intentional right action. Doing the right thing is a matter of complying with a moral norm. In this respect, doing the right thing is different from, for example, basic action – it's different from lifting your arm or furrowing your brow. These things we can do directly. In contrast, complying with a norm is something that we do *by* performing some (more basic) action. We conform to a norm by performing an action that complies with the norm. Thus, Superman does the right thing *by* saving the children from inside a burning building and thereby complying with the moral norm to help others in need.

To intentionally save the children, Superman must know how to save the children. But this is not enough to intentionally do the right thing. Superman is not intentionally acting rightly if he is saving the children from the burning building for some ulterior purpose – for example, because he plans to drown them. But he's also not intentionally acting rightly if he has absolutely no idea that saving the children is the right thing to do or if he mistakenly takes it to be the wrong thing to do. In either case, Superman is not intentionally doing the right thing because he does not know how to do the right thing.

Intentionally doing the right thing, thus, requires moral knowledge: it requires an agent to know what the right thing to do is. An agent intentionally does the right thing only if she knows what the right thing to do is.

3. Why Moral Praise requires Intentionally Doing what's Right

What the agent did intentionally matters to our moral evaluation of the action. Just consider Kant's shopkeeper, who does not overcharge his customers but not because being honest is the right thing to do – rather it's because he thinks that honesty is most likely to keep his business profitable. Or consider the CEO who decides on a policy that will benefit the environment – but only because the policy also happens to be good for the bottom line. In both cases, we are not willing to give the agent's moral credit for their right action. And it's natural to justify this reluctance by appealing to the fact that their doing the right thing was not intentional. This is not to deny that there are some actions that the two agents intentionally performed. Presumably, the shopkeeper did intentionally charge each customer the relevant price for the goods. The CEO did intentionally adopt the relevant policy. Nevertheless, they didn't intentionally do the right thing: their actions were not guided by an intention to do what's right along with knowledge of what the right thing to do is.[7]

Our judgments about these cases are suggestive but there is a difficulty. You might question whether these cases can lend support to my claim that moral praise requires moral knowledge. After all, it looks like the real work here is being done by

---

[6] See Williamson [2000], chapter 1 for a defense that knowledge is unanalyzable. An alternative conclusion is that intentional action is primitive. See Levy [2014] for a proposal along those lines.

[7] My account of intentionally doing what's right is compatible with the experimental results reported by Knobe [2003a]. For an overview, see Knobe and Pettit [2009]. See, Holton [2010].

the agents' intentions. Ultimately, it's because the CEO and the Kantian shopkeeper do not intend to do the right thing that they fail to be morally praiseworthy for doing the right thing. Perhaps then what matters to moral praise is not whether the action is intentional or not but whether the agent has the right intentions. Lack of such an intention makes the agent's doing the right thing unintentional – but it doesn't yet show that intentionally doing the right thing is necessary for moral praiseworthiness. It's still a possibility that an agent could be morally praiseworthy for doing the right thing unintentionally when they act unintentionally solely because of a lack of moral knowledge.

There are two reasons for thinking that moral knowledge does matter to our moral assessments. First, we can reflect on our practice of giving people moral credit for their actions. As Watson has noted, we often remark on praiseworthy actions by saying things like "that was good of you/her," or "she really deserves credit for doing the right thing there." Note just how very odd it would be to remark upon someone's doing the right thing: "She really deserves credit for doing the right thing there. Of course, she had no idea what the right thing to do was in that situation!" "It was really good of her to help here; it was the right thing to do. Of course she had no idea how to do the right thing in that situation!" These remarks have an air of paradox – so much so, that it's hard not to read them as intended to express sarcasm. If, as I have suggested, our practice of giving moral credit for right actions tracks whether the agent's doing the right thing was intentional, this is hardly a surprise: in saying that the agent deserves moral credit for her action, we imply that we take her to have done the right thing intentionally. The claim that the agent lacked knowledge how to do the right thing directly contradicts this.

Second, moral praise is incompatible with certain kinds of moral luck. To be morally praiseworthy for an action, her having done the right thing needs to be counterfactually robust. It cannot just be an accident that she has acted rightly. As Kant puts it:

> In the case of what is to be morally good it is not enough that it [i.e. the action] conform with the moral law but it must also be done for the sake of the law; without this that conformity is only very contingent and precarious, since a ground that is not moral will indeed now and then produce actions in conformity with the law, but it will also often produce actions contrary to law.[8]

But the mere fact that a right action was produced by an intention to do the right thing is not enough to guarantee such counterfactual robustness; the proverbial road to hell is paved with good intentions. Good intentions will only lead to right actions non-accidentally, if they are paired with knowledge of what the right thing to do is.[9] But if, as I have argued, our attributions of praise track whether the agent intentionally acted rightly, we have an explanation for why we expect morally worthy actions to be counterfactually robust in certain ways. Morally worthy actions inherit this counterfactual robustness from the counterfactual robustness of intentional action; which in turn, inherits it from its epistemic condition.

You might question whether moral worth really does require counterfactual robustness of this sort. Suppose that my friend has a sudden heart attack. I don't know how to perform CPR but there is no time to loose; if I do nothing, my friend will die. I perform the motions that I vaguely remember having seen on TV, guessing how hard and how often to press down on his chest. My guess was in the right

---

[8] Kant (4:390).

[9] For an in-depth development of this argument, see Sliwa [2015].

ballpark; my friend survives. Don't I deserve moral credit? Of course; but the question is which of my actions I deserve credit for. Performing CPR was the right thing to do. But, crucially, in a situation where I do not know how to perform CPR, *attempting* to perform CPR is also the right thing to do. And while I did not know how to do the former, I did know how to do the latter. As it happens my attempt was successful and I saved my friend's life. But I would have deserved moral credit for the attempt even if I had made the wrong guess. And so while it is lucky that I managed to perform CPR; it's not lucky that I succeeded in doing the right thing. My attempt at CPR was the right thing to do and it was guided by my knowledge how to do what's right.

Let's take stock. This section was concerned with setting out some important motivations for the claim that moral worth requires intentionally doing what's right. Next, I will focus on what I take to be the most controversial aspect of this claim: since intentionally doing what's right involves being guided by one's moral knowledge, moral worth requires the agent to conceptualize her action as the right thing to do. The next section defends this claim against some apparent counterexamples that have been prominent in the literature. I will then make a positive case in support of it.

4. Moral Conceptualization and Intentional Right Action

One central criticism of Premise 2 is that what it requires for moral worth is too demanding: it fails to give moral credit to many agents to whom, intuitively, moral credit is due. In particular, it implies that to deserve moral credit for her action, the agent must conceive of what she's doing as the right thing to do.

Arpaly & Schroeder argue that agents often act in morally admirable ways despite failing to conceive of what they are doing as morally right. They put forward the following example:

> ...imagine an undergraduate student, Brandon, whose moral view (greatly influenced by the writings of Ayn Rand) is that one should be selfish. Not just that selfish behavior is his moral right, but that it is his "sacred," as he would say, "moral duty." Nonetheless, Brandon often acts unselfishly. Typically he just fails to notice his failure to conform to his theoretical standards; occasionally he berates himself for his "sentimentality" when he sees that he is contributing to "weak, degenerate, socialistic" practices rather than acting selfishly and so "getting something out of" what he is doing."[10]

Arpaly & Schroeder argue that because Brandon's subscribes to false moral principles, he is not in a position to conceive of his unselfish acts as the right thing to do. Nevertheless, they argue, we should give him moral credit when, for example, he helps his classmate: while "Brandon's Ayn Rand-centered beliefs show that he is a bad philosopher", being a bad philosopher is fully compatible with acting in morally praiseworthy ways.[11]

I agree that acting in a morally admirable way is fully compatible with being a bad philosopher; but this is because being a bad philosopher is compatible with intentionally acting rightly. Generally agents who are bad moral philosophers and decent people are not fully in the grip of their false moral principles; they only *kind of* believe them, in *some* contexts (such as the classroom, or when engaging in abstract moral reflection) but not others (when confronted with someone needing help). Drawing on recent philosophy of mind, we can characterize their beliefs are

[10] Arpaly & Schroeder [2014], p. 177.

[11] ibid.

fragmented.[12] They are irrational, they have incoherent beliefs. But their failing to know how to do what's right in abstract moral deliberation is compatible with their knowing how to do what's right in many other circumstances.

And so, I am also inclined to agree that Brandon deserves moral credit for his unselfish actions – at least if we fill in the details of the case in a very plausible way. There are some contexts in which Brandon believes that he ought to be selfish – in particular, contexts in which he is abstractly reflecting on what to do. But in most other contexts, Brandon knows that he ought *not* be selfish: that he should help his peers, tell the truth, and keep his promises. And this knowledge guides his actions. Thus, when Brandon helps his peer to pick up his papers, it's not just an accident that he acted rightly. He acted rightly because he wanted to do what's right and he knew how to do that: his right action was intentional.

I have argued that, contrary to Arpaly & Schroeder, agents like Brandon typically do conceive of helping their friend as the right thing to do – their misguided moral principles none-withstanding. They do intentionally act rightly. But what about an agent who unintentionally acts rightly because she mistakenly judges her action to be morally wrong? The central example here is the much discussed case of Huckleberry Finn. Huckleberry helps the fugitive slave Jim by making up an elaborate lie and thereby protecting him from being captured. But, Huckleberry suffers from what Arpaly calls "inverse akrasia": he believes that he should tell on Jim and return him to his "rightful owner" Miss Watson. Since Huckleberry believes that he is acting wrongly, he is not intentionally acting rightly in protecting Jim. Nevertheless, so the received wisdom, he deserves credit.[13]

I agree that Huckleberry deserves some credit; but it's not clear that he deserves *moral* credit. Moral standards are not the only standards against which we evaluate the actions of others. We praise others for being good friends, good doctors, or good department citizens. Huckleberry has false beliefs about what what morality requires of him in the particular situation he finds himself in and as a consequence he does not know how to do the right thing in the situation he finds himself in. But he plausibly does know how to be a good friend. Being a good friend is something that we value and admire for its own sake; and so Huckleberry's action rightly strikes us as meriting praise. However, the norms of friendship are not the norms of morality; being a good friend can sometimes require you to do something that is morally wrong.[14] The present account also need not deny that Huckleberry is a good boy; that he can and often does intentionally act rightly.[15] But this is compatible with him not being morally praiseworthy for the particular action of helping Jim.

5. In Defense of Moral Conceptualization

The fact that the present view has plausible things to say in response to these counterexamples gives us a reason not to dismiss it. But it's not yet a reason to prefer it over alternatives that do not require the agent to conceptualize her action as right. This section aims to make a positive case in defense of this requirement, arguing that we should prefer the present account of moral worth to a prominent competitor put forward by Arpaly & Schroeder: Spare Conativism.

---

[12] See, in particular Schwitzgebel [2010], Elga & Rayo [ms.] and Marley-Payne [ms.]

[13] There is widespread agreement on this verdict in the literature. See, for example, Arpaly [2003], Markovits [2010], Driver [2001].

[14] See Cocking & Kennett [2000].

[15] Arpaly [2003], p. 73.

According to Spare Conativism, an action is morally praiseworthy when it is rationalized by an agent's good will. Good will, in turn, is a matter of having desires with the right content and of the right kind. First, the desires need to concern to what is in fact the right to do, in terms of the concepts of the correct normative theory. Thus, insofar as, for example, Act-Consequentialism is the correct normative theory, being good consists in having desires to perform those actions that in fact maximize utility conceived *qua* actions that maximize utility. Second, these desires need to be intrinsic, not merely instrumental. Thus, Arpaly & Schroeder argue:

> The reference of an intrinsic desire that counts as complete good will must, naturally, be the right or the good. But [...] a given referent can be conceptualised in many different ways. Spare conativism holds that the sense required for perfect good will is to be determined by normative moral theory: the concepts deployed in grasping the correct normative moral theory are the concepts through which one must intrinsically desire the right or good in order to have good will.[16]

On Spare Conativism then, an agent does not need to intentionally do the right thing in order to deserve moral credit for the action because she does not need to conceptualize her action as the right thing to do. She does, however, need to conceptualize her action in terms of the correct normative theory. The main disagreement between the view defended here and Spare Conativism thus concerns *which* intentional action an agent needs to perform. According to Spare Conativism (and assuming again that Act-Consequentialism is the correct normative theory), the agent needs to intentionally maximize happiness; she does not need to intentionally do the right thing.[17]

I think Spare Conativism is mistaken; conceptualizing one's action as the right thing to do matters both for having good will and for one's actions having moral worth. To see why, we need a better grip on what such conceptualization involves. Concept possession is, in part, a matter of categorization. Think about what's involved in someone seeing a red object as a red object. It's a matter of exercising a certain discriminatory ability: the ability to discriminate red objects from those that are not red, as well as to categorize other red objects as relevantly similar to the one observed. Similarly, conceptualizing an action (such as giving up one's seat) as right requires the agent to "see it" as belonging to the same category as other right actions (e.g. keeping a promise, asking for someone's consent) and as different from both actions that are wrong (breaking a promise, pushing someone) and those that are merely required by social norms (using the outermost set of cutlery for the appetizer).

Conceptualizing something *as right,* however, involves more than just categorization. Normative concepts, and moral concepts in particular, are special in that they play a distinct practical role. As Kahane argues:

> Our evaluative discourse plays a certain role in our practical lives. [...] What is this role? It's not easy to spell it out in entirely neutral terms, but the basic idea is simple: it's the role of setting a standard by which attitude and action can be made

---

[16] Arpaly & Schroeder [2014], p. 164.

[17] Arpaly & Schroeder [2014] do not frame their accounts explicitly in terms of intentional action. But they hold that in order to act on the right reasons, the agent's action needs to be rationalized by the right kind of belief-desire pair. Insofar as only intentional actions admit of such rationalizing explanations (see, e.g. Davidson [1963] for the classical treatment of this), they too are committed to the principle that an agent can be morally praiseworthy only for what she does intentionally.

intelligible and justified, and in light of which we deliberate (in the first-person), and give advice or criticize (in the second- and third-person).[18]

To conceptualize something as right or wrong is thus to regard it as a standard against which actions (one's own and those of others') can be measured against. For the purposes of this paper, I want to focus on one particular aspect of this practical role, namely the link between moral concepts and criticism. A plausible way to spell out this connection is in terms of reactive attitudes such as blame, guilt, remorse, indignation, gratitude, admiration. Thus, to conceiving of something as the right (or wrong) thing to do involves the disposition to experience certain reactive attitudes. Second, it involves seeing these reactive attitudes *as appropriate*. Thus, to conceive of something as morally wrong is to take it to be the kind of action that warrants a particular kind of criticism: blame by others and remorse by oneself. To see something as morally right is to see it as the kind of action such performing is admirable and that failing to perform it makes one the legitimate target of a particular kind of criticism by others: blame, resentment.

It will also be helpful to put on the table what conceptualizing something as the right thing to do need *not* involve: first, it need not involve referring to it by the word "right". (Just as conceiving of something as chocolate is not just a matter of using the word "chocolate" to refer to it. Someone who calls all brown things "chocolate" does not thereby have the concept of chocolate.) [19] Words generally express concepts but one can have a particular concept without being able to express it. Second, conceptualizing something as the right thing to do is not a matter of thinking about morality or engaging in moral reflection – just as to conceptualize something as red does not require that the agent think or deliberate about colors. For this reason, an agent who conceives of something as the right thing to do need not be guilty of "one thought too many": the exercise of her discriminatory ability need not be conscious.

While this is but a sketch, it's enough to see that Spare Conativism is missing out on a crucially important aspect of moral agency. Take an agent whose intrinsical desires are perfectly aligned with Act-Consequentialism, which is (let's assume) the correct normative theory. She has an extremely strong desire to maximize happiness and minimize suffering. But, let's also assume, she does not conceive of maximizing happiness or minimizing suffering *as* the right thing to do. And when she acts in light of her intrinsic desires, she does not conceptualize her actions as right. Such an agent would, of course, reliably do what is in fact right. Nevertheless, there is something very odd and disturbing about her. While she strongly desires to maximize happiness, she does not think of it as a standard for action. And this means that she does not see failures to comply with this standard as meriting a particular kind of response on the part of herself and others: of blame, guilt, remorse. She may, of course, be very frustrated when she herself or someone else fails to maximize happiness. But this frustration is simply that of not having others act as one wants them to; it does not differ in kind from the frustration that one might experience

---

[18] Kahane [2013], p. 157. See also Eklund [2012]. The point here also echoes McDowell [1979]:

> A kind person can be relied on to behave kindly when that is what the situation requires. Moreover, his reliably kind behaviour is not the outcome of a blind, non-rational habit or instinct, like the courageous behaviour – so called only by courtesy – of a lioness defending her cubs. [...]A kind person has a reliable sensitivity to a certain sort of requirement that situations impose on behaviour. The deliverances of a reliable sensitivity are cases of knowledge; and there are idioms according to which the sensitivity itself can appropriately be described as knowledge: a kind person knows what it is like to be confronted with a requirement of kindness. (p. 51)

[19] Foot [1978], p. 120

when one is stuck in traffic behind a slow drive or when one is forced to listen to a neighbours' playing the Titanic song ("My heart will go onnnnn aaaaand onnnnn") on repeat.

This suggests that there is more to good will than just having a set of intrinsic desires that will reliably produce actions in accordance with the moral requirements. Good will also involves being disposed to have the appropriate responses to those actions that meet or fail to meet these requirements: to feel guilt and remorse or blame and indignation rather than mere frustration that one's desires have not been satisfied. And this, in turn, is a matter of conceiving of these actions in moral terms; of assessing them with regards to a moral standard.

I believe that Arpaly & Schroeder are also mistaken when they say that it's possible to be morally praiseworthy for what one does even as one conceptualizes it as the wrong thing to do; and hence, that intentionally doing what's right is not required for morally praiseworthy action. Their case rests on our intuitions about the admirableness of Huckleberry Finn. But, as I have argued above, it's not clear that our intuitions about Huckleberry Finn genuinely track the moral admirableness of his *action* – we might be giving Huckleberry nonmoral credit for being a good friend or we might be responding to the fact that Huckleberry is just generally likable.

One way to adjudicate between these competing explanations is to consider a different case of unintentionally right action by someone who is neither moved by friendship nor as likable as Huckleberry. If Arpaly & Schroeder are right that intentionally doing the right thing really is not necessary for moral worth, this action should strike us as unambiguously morally admirable. But I do not think that it does. Consider the case of a Stalinist party activist who participates in Stalin's program terror-famine in the Ukraine. This involved systematically searching the peasant's houses for food and confiscating it, to break the resistance of the "rich peasants" to Stalin's plans of farm collectivization.The party activist acknowledges that the means to this end is gruesome and that he often finds his work tough – he has to collect all food down to the last bite, closing his ears to the desperate petitions of parents and the cries of their emaciated children. But he ardently believes in the final cause and he takes any hesitation or moral doubt to be a sign of weakness and "intellectual squeamishness".[20] Suppose that on one of his searches, he finds and takes away a meagre portion of food from a family. And just as he is about to leave, he notices, from the corner of his eye, that their little son seems to be clutching something in his little fist that might be a morsel of bread crust. He "knows" that it is "his duty" to confiscate everything but in this particular case, he just can't bring himself to do it; for a fleeting moment the figure reminds him of his own child. He willfully overlooks the child all the while scolding himself for his weakness and sentimentality. He feels shame for his "moral weakness" and for the next couple of days he grimly doubles-down on his search mission, executing it with more fervor and fastidiousness than ever before.

There are two differences here to Huckleberry Finn's case. First, the party activist is a rather unsavory character. Second, it's much clearer that he conceptualizes his right action as the wrong thing to do. On Arpaly & Schroeder's account, however, neither of these should bear on whether the actual action he performs is morally admirable. And as far as his action is concerned, he both acted rightly in letting the child keep his bite of food and his action seems to be rationalized by the kind of intrinsic desires that Arpaly & Schroeder take to suffice for moral worth: he was,

---

[20] My exposition of the case follows closely that of a real case, described by Baumeister [1997], p. 197-80.

after all, moved by a sense of the child's humanity.[21] But while we may be glad that he experienced a moment of "weakness", I do not think that we are inclined to give him moral credit for it. What's missing is any recognition that the party activist is not just acting on a whim but that he's doing something that he *should* do. His certainty that he is acting wrongly does not waver for a moment; his attitude to his own action is solely one of remorse, guilt, and disappointment. In short, the problem is that he genuinely conceives of what he has done as the wrong thing to do.

6. Intentionally Doing What's Right and the Limits of Moral Responsibility

My focus so far has been on the conditions when agents are morally praiseworthy for an action. This is a question about when it's appropriate for us to adopt a particular kind of reactive attitude towards the agent: gratitude, admiration, giving her credit. I have argued that it's appropriate to take this kind of attitude when the agent did the right thing, intentionally. To adopt a particular reactive attitude towards an agent in response to her action is to hold the agent responsible for that action.

I want to end by turning to a broader question. Just as we can ask about when it's appropriate to take a particular reactive *attitude*, we can also ask about when it's appropriate to take the reactive *stance* towards an agent. To take the reactive stance towards someone is not just a matter of holding them morally responsible for some particular action. Rather, it's a matter of ascribing to them moral agency – the kind of agency that goes hand in hand with being a morally responsible agent. To say that the reactive stance towards an agent is unwarranted is thus to say that the other person is exempt from attributions of moral responsibility. A person who is exempt from moral responsibility may still be an agent in a causal sense; just as an animal or a small child. But they are not a moral agent in the sense that they are not the legitimate object of a whole range of reactive attitudes. As Strawson puts it, we can see such a person as:

> an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained; perhaps simply to be avoided, though this gerundive is not peculiar to cases of objectivity of attitude.[22]

What does the view defended here tell us about when it's appropriate to forgo the reactive stance in favour of the objective stance? As I have argued, what makes it appropriate to adopt reactive attitudes of praise, admiration, gratitude is the other person's intentionally having acted rightly. Thus, I suggest, that it is appropriate to adopt the reactive *stance* towards an agent if she has the *ability* to perform morally praiseworthy actions. Since moral worth requires intentionally acting rightly, we can pin down the specific capacities involved; it's those capacities that matter for intentionally doing the right thing. For one, this involves a general capacity to act intentionally – a capacity that is not specific to moral agency. The agent must be capable of planning, of forming intentions, and acting in light of them. The second is an epistemic capacity. To intentionally act rightly, the agent must know what the right thing to do is. This means that intentional right action requires the capacity for moral knowledge; it requires moral competence – the ability discriminate right from wrong.

---

[21] To put it in Arpaly & Schroeder's terminology his action manifests partial "reverse moral indifference".

[22] Strawson [1962], p. 10

Not all impairments of moral agency call for a suspension of reactive attitudes. For example, discussing the case of mass murderer Robert Harris, Watson notes:

> To be homicidally hateful and callous [...] is to lack moral concern, and to lack moral concern is to be incapacitated for moral community.[23]

But the mere fact that someone is "homicidally hateful and callous" is hardly a reason to suspend our reactive attitudes. It's not the kind of consideration that gives an agent a "free pass" when it comes to moral responsibility. On the other hand, other conditions do make it appropriate to take the objective stance towards the agent: such as when she is in the grip of psychosis, or severely depressed, or having suffered from certain forms of brain damage. A plausible account of moral responsibility must differentiate between those impairments of moral agency that exempt an agent from moral responsibility attributions and those impairments that make her bad. One strength of the present account is that it meets this challenge.

In contrast, this challenge that makes trouble for competing accounts of moral responsibility. Consider again Spare Conativism. According to Spare Conativism the sole determinant of the agent's moral goodness (and correspondingly, her moral badness) are her intrinsic desires. Good will is a matter of having an intrinsic desire for the right or good and ill will is an intrinsic desire for the wrong or bad – correctly conceptualized. Just as ill will is a vice, so is moral indifference: the lack of intrinsic desires for the right and good. Spare Conativism does not exempt those who are "homicidally hateful and callous" from moral responsibility. After all, being "homicidally hateful and callous" plausibly just involves the kinds of intrinsic desires that constitute ill will. But Spare Conativism faces the opposite problem: it's too quick to count people as vicious – either as having ill will or as being morally indifferent – whose moral agency is impaired. Many of the circumstances that we generally regard as exempting conditions *just are* conditions that influence which desires an agent has – either directly or indirectly.

Consider, for example, a woman suffering from severe postnatal depression. Postnatal depression has many manifestation but in it often involves an inability to bond with one's child. In its more severe forms, it may involve anger and resentment towards one's child – even a desire to harm the child. Postnatal depression thus may directly impair a woman's concern for an important right-making reason (her child's wellbeing) and it may even involve her coming to have a "sinister" desire to harm her child. On the fact of it then, postnatal depression gives an agent (partial) ill will. But this seems implausible. Severe postnatal depression does not make someone *evil*; it calls for treatment, not for moral condemnation.

A second problem arises from conditions which impair an agent's intellectual ability to grasp the concepts that figure in the correct normative theory. A grasp of these concepts is needed in order to have the intrinsic desires that, on Arpaly & Schroeder's account, constitute good will.[24] But which concepts we are in a position to grasp depends, in part, on our epistemic and cognitive faculties. This gives rise to two worries. The first is that the required concepts may well be fairly complex. Just think about the concepts that are commonplace in contemporary ethical theories: the concept of rights, of informed concept, of autonomy. These are theoretical notions that differ in significant ways from any folk notions in the vicinity. Their grasp requires knowledge of the theory itself as well as fairly sophisticated cognitive abilities. These might well be beyond the reach of agents with intellectual disabilities.

---

[23] Watson [2002], p. 242.

[24] These worries are somewhat more speculative, since Arpaly & Schroeder do not commit themselves to any particular normative theory.

Second, consider an agent with a neurological condition that impairs her theory of mind. Insofar as the correct normative theory involves concepts that make reference to other's mental or emotional state (such as certain notions of well-being), such an agent, too, might be precluded from grasping them. But having an intellectual disability or a neurological condition does not make one *morally indifferent*.

On the present view not all impairments of moral agency license an adoption of the objective stance. In particular, it does not exempt those who are "homocidally hateful and callous". This is because being "homocidally hateful and callous" is both compatible with having the ability to form intentions and act in light of them as well as with knowing how to do what's right. An agent who systematically does what she knows to be wrong, simply because she does not care about doing what's right or because one because she cares more about other things – money, revenge, status – is morally blameworthy. Lack of moral commitment is not an exempting condition. It's compatible with full moral agency.

But, unlike Spare Conativism, it does have the resources to exempt agents with severe postnatal depression, psychosis, certain cognitive and neurological disorders. This is because these conditions often do affect the capacities that, on the present account, lie at the heart of moral agency: they impair the agent's general capacity for intentional action. Or they systematically limit her capacity to discriminate right from wrong (or both). And so, they drastically limit an agent's ability to intentionally act rightly.

This gives us a principled answer as to what constitutes exempting conditions for moral responsibility and why. What exempts an agent from moral responsibility are conditions that drastically impair either her intentional agency – i.e. her capacity to form intentions and act in light of them – or her moral competence – her capacity to discriminate right from wrong – or both of those. These are exempting conditions because moral agency requires the agent requires the ability to intentionally do the right thing. An agent who lacks one of these capacities thus lacks moral agency and an agent whose capacities are greatly diminished is one whose moral agency is greatly diminished.[25]

Finally, the present account makes sense of why learning about an agent's history can influence whether we take the reactive stance towards them. This addresses a central worry raised by Watson, in the context of his discussion of Robert Harris. Watson observes that while we are willing to condemn Harris for his heinous crimes, our attitudes shift once we learn about his sad and deprived childhood. But what is it about someone's deprived childhood that justifies letting them off the moral hook?

> Does Harris have some independently identifiable incapacity for which his biography provides evidence?...To be homicidally hateful and callous in Harris's way is to lack moral concern, and to lack moral concern is to be incapacitated for moral community. However, to exempt Harris on these grounds is problematic. For then everyone who is evil in Harris's way will be exempt, independently of facts about their background. But we had ample

[25] The view I have arrived at has important similarities to that defended by Susan Wolf. In particular, it agrees that *sanity* is a necessary condition for moral responsibility. Wolf [1990], p. 77 argues:

> [T]he crucial feature that distinguishes responsible beings from others...according to the Reason View, is the ability to be in touch with the True and the Good. In other words, what makes responsible beings special is their ability to recognize good values as opposed to bad ones and to act in a way that expresses appreciation of this recognition. The freedom and power necessary for responsibility, then, are the freedom and power to be good, that is, the freedom and power to do the right thing for the right reasons.

evidence about this incapacity before we learned of his childhood misfortunes, and that did not affect the reactive attitudes.[26]

Watson is right that a systematic lack of moral concern is *not* enough to exempt one from responsibility for one's wrong actions. But in contrast to Watson, I do think that learning about Harris' childhood does provide us with evidence about an "independently identifiable incapacity". In his narrative of Harris' crimes, Watson presents Harris as someone who intentionally acts wrongly: someone who deliberately seeks out wrong actions because they are wrong.[27] Learning about the constant cruelty he suffered rightly gives us pause because it casts doubt on the central presuppositions of this narrative. It calls into question Harris' moral competence. How could a child subjected to such relentless abuse develop a sense of right and wrong?[28]

Bibliography

Anscombe, G. E. M. (1957). *Intention*. Harvard University Press.

Arpaly, Nomy (2003). Unprincipled Virtue: An Inquiry Into Moral Agency. Oxford University Press.

Arpaly, Nomy and Schroeder, Timothy (2014). *In Praise of Desire*. Oxford University Press.

Baumeister, Roy (1997). Evil: Inside Human Cruelty and Violence. New York: W. H. Freeman and Co.

Bratman, Michael (1987). *Intention, Plans, and Practical Reason*. Center for the Study of Language and Information.

Cocking, Dean & Kennett, Jeanette (2000). Friendship and moral danger. *Journal of Philosophy* 97 (5):278-296.

Davidson, Donald (1963). Actions, reasons, and causes. *Journal of Philosophy* 60 (23): 685-700.

Driver, Julia (2001). *Uneasy Virtue*. Cambridge University Press.

Eklund, Matti (2012). Alternative Normative Concepts. *Analytic Philosophy* 53: 139-57.

Elga, Adam and Rayo, Agustin (ms.) Fragmentation and Information Access.

---

[26] Watson, p. 242-43.

[27] For example: "Unlike the small child, or in a different way the psychopath, he [i.e. Harris] exhibits an inversion of moral concern, not a lack of understanding. His ears are not deaf, but his heart is frozen." (p.239)

[28] There is evidence that psychopathy involves serious deficits in moral competence. In particular, psychopaths seem to have difficulties discriminating between conventional and moral norms. For an overview, see Kennett & Fine [2008]. But the ability to draw such a distinction seems central for a grasp of moral concepts. And so these deficits go against the received wisdom that psychopaths know perfectly well right from wrong; they just don't care.

Fine, Cordelia & Kennett, Jeanette (2004). Mental impairment, moral understanding and criminal responsibility: Psychopathy and the purposes of punishment. *International Journal of Law and Psychiatry* 27 (5):425-443.

Gibbons, John (2001). Knowledge in action. *Philosophy and Phenomenological Research* 62 (3):579-600.

Glick, Ephraim (forthcoming). Abilities and Know-How Attributions. In Jessica Brown & Mikkel Gerken (eds.), *New Essays on Knowledge Ascriptions*. OUP.

Hawley, Katherine (2003). Success and Knowledge-How. *American Philosophical Quarterly* 40 (1):19 - 31.

Herman, Barbara (1981). On the value of acting from the motive of duty. *Philosophical Review* 90 (3):359-382.

Holton, Richard (2010). Norms and the Knobe Effect. *Analysis* 70 (3):1-8.

Kahane, Guy (2013). Must Metaethical Realism Make a Semantic Claim? *Journal of Moral Philosophy* 10 (2):148-178.

Kant, Immanuel (1996). *Practical Philosophy*. Cambridge University Press.

Knobe, Joshua (2003a). Intentional action and side effects in ordinary language. *Analysis* 63 (3):190–194.

Knobe, Joshua (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology* 16 (2):309-325.

Pettit, Dean & Knobe, Joshua (2009). The Pervasive Impact of Moral Judgment. *Mind and Language* 24 (5):586-604.

Levy, Neil (2014). Consciousness and Moral Responsibility. Oup Oxford.

Levy, Yair (2013). Intentional action first. *Australasian Journal of Philosophy* 91 (4): 705-718.

Markovits, Julia (2010). Acting for the right reasons. *Philosophical Review* 119 (2): 201-242.

Markovits, Julia (2012). Saints, heroes, sages, and villains. *Philosophical Studies* 158 (2):289-311.

Marley-Payne, Jack (ms). Task-Indexed Belief.

McDowell, John (1979). Virtue and Reason. *The Monist* 62 (3):331-350.

Mele, Alfred R. (1992). *Springs of Action: Understanding Intentional Behavior*. OUP.

Mele, Alfred R. & Moser, Paul K. (1994). Intentional action. *Noûs* 28 (1):39-68.

Schwitzgebel, Eric (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly* 91 (4):531-553.

Setiya, Kieran (2012). Knowing How. Proceedings of the Aristotelian Society 112 (3):285-307.

Setiya, Kieran (2008). Practical knowledge. *Ethics* 118 (3):388-409.

Setiya, Kieran (2003). Explaining action. *Philosophical Review* 112 (3):339-393.

Sliwa, Paulina (forthcoming). Moral Knowledge and Moral Worth. *Philosophy and Phenomenological Research*.

Strawson, Peter F. (1962). Freedom and resentment. *Proceedings of the British Academy* 48:1-25.

Velleman, David (2000). The Possibility of Practical Reason. Oxford University Press.

Watson, Gary (2004). Agency and Answerability: Selected Essays. Oxford University Press.

Williamson, Timothy (2000). *Knowledge and its Limits*. Oxford University Press.

Wolf, Susan (1990). *Freedom Within Reason*. Oxford University Press.